УДК 68Т50

А.О. Никоненко
Київський національний університет імені Тараса Шевченка, Україна
вул. Володимирська, 64/13, м. Київ, 01601

МЕТОД ВИЗНАЧЕННЯ СЕМАНТИЧНОЇ ЗВ'ЯЗНОСТІ

A.O. Nykonenko
Taras Shevchenko National University of Kyiv, Ukraine
Volodymyrska st., 64/13, Kyiv, 01601

SEMANTIC RELATEDNESS CALCULATION METHOD

Роботу присвячено вивченню проблеми визначення семантичної зв'язності понять англійської мови на базі текстових корпусів. На початку роботи ми наводимо короткий огляд існуючих підходів до вирішення проблеми, розглядаємо основні еталонні корпуси, що розмічено експертами. Далі переходимо до опису власного методу та основних класів гіпотез, на яких він базується. В роботі запропоновано і описано більше 70 гіпотез, що можуть бути використаними при обчисленні семантичної зв'язності, а також нову, високоефективну модель вимірювання зв'язності на базі машинного навчання і запропонованих гіпотез. Модель дозволяє гнучко обирати серед гіпотез підмножини і показує високу ефективність на різних наборах еталонних тестів.

Ключові слова: семантична зв'язність, дистрибутивна семантика, машинне навчання.

The work is dedicated to the problem of semantic relatedness calculation based on text corpora. At the beginning of the work, we present a brief overview of existing approaches to solve the problem and consider the basic benchmark corpora. Then we describe our own method and main hypotheses on which it is based. The paper presents more than 70 hypotheses that can be used in the calculation of semantic relatedness and a new, high-performance relatedness measure model based on machine learning. The model can flexibly switch between subsets of hypotheses and demonstrate high efficiency on different benchmarks sets.

Keywords: semantic relatedness, distributional semantics, machine learning.

Вступ

В питанні вимірювання ступеню схожості слів існує три досить близьких, проте не синонімічних поняття: семантична схожість (semantic similarity), семантична зв'язність (semantic relatedness) і семантична відстань (semantic distance). Дана робота присвячена розробці нового методу вимірювання семантичної зв'язності. В огляді [1] стверджується, що поняття семантичної зв'язності є більш загальним чим поняття семантичної схожості, оскільки, з точки зору онтологічної бази знань, включає набагато більшу множину зв'язків між поняттями, в тому числі антонімію, меронімію та інші. Автори стверджують, що вирішення лінгвістичних задач в більшості випадків потребує обчислення саме семантичної зв'язності, а не більш вузької семантичної схожості. Наприклад, у вирішенні задачі визначення смислу поняття (word sense disambiguation) можуть використовуватися будь-які зв'язки між словом і контекстом, а не лише зв'язки з близькими поняттями.

Моделі вимірювання схожості-зв'язності

Існують два базові підходи до підрахунку семантичної схожості-зв'язності: топологічний і статистичний. Топологічний включає в себе множину методів, які в якості джерела даних про мову використовують структуровані ресурси: словники, онтології, Вікіпедію і т.д. Статистичний підхід базується на використанні корпусів текстів та зібраних з них статистик про частоту сумісної зустрічаємості слів в рамках певного контексту (вікна в N слів, речення, абзацу, документа). На протязі кінця 90-х – початку 00-х бурхливого розвитку зазнали топологічні методи, в першу чергу це було пов'язано з популяризацією таких проектів, як WordNet [2] і CYC [3] та

експотенційного росту числа статей у Вікіпедії. Останню декаду набирає популярності статистичний підхід, що пов'язано з доступністю та відносною легкістю створення текстових корпусів. Моделі, що працюють на базі другого підходу, називаються моделями дистрибутивної семантики.

Наразі існує велика кількість моделей дистрибутивної семантики [4], що відрізняються лише параметрами:

• тип контексту: розмір контексту, правий чи лівий контекст, ранжування;

• кількісна оцінка частоти слів в контексті: абсолютна частота, TF-IDF, ентропія, сумісна інформація і т.д.

• міра відстані між векторами: косинусна, скалярний добуток, відстань Мінковського, Евклідова, Манхеттенська і т.д.

• методи зменшення розмірностей матриць: випадкова проекція, сингулярний розклад, випадкове індексування і т.д.

На базі даних моделей створено велику кількість методів вимірювання семантичної зв'язності, з якими ми будемо порівнювати точність роботи нашого методу.

Оцінка точності методів

Вимірювання точності методів схожості-зв'язності відбувається на спеціальних розмічених людьми корпусах. Більшість таких корпусів створюється для вимірювання семантичної схожості і наразі існує всього чотири корпуси для вимірювання семантичної зв'язності:

• WS353-R[5] – автори роботи вручну розбили WordSim353 на два набори даних: WS353-S, що містить 203 пари схожих понять, та WS353-R, що містить 252 пари зв'язаних понять. Кожна пара розмічена за шкалою 0-10.

• REL-122[6] – один з нових наборів даних, містить 122 пари слів, розмічених від 0 (незв'язані поняття) до 4 (сильно зв'язані), кожна пару анотували від 14-ти до 22-х чоловік.

• Mturk-287[7] – набір містить 287 пар слів, кожна пара розмічена за шкалою 1-5 десятьма людьми.

• Mturk-771[8] – набір містить 771 пару слів, кожна пара розмічена за шкалою 1-5 двадцятьма людьми.

В рамках даної роботи ми нормалізуємо значення оцінок всіх еталонних корпусів (зводимо до інтервалу [0,1]) для простоти подальшого порівняння отриманого результату з еталоном.

Для оцінки точності роботи методів оцінки зв'язності на перерахованих наборах даних використовується взаємна кореляція. Для числового вираження ступеню взаємної кореляції в даній задачі використовується коефіцієнт кореляції Спірмена [9] та коефіцієнт кореляції Пірсона[10]. Через певні особливості функціонування коефіцієнту Пірсона дослідники віддають перевагу коефіцієнту Спірмена, як більш стабільному. В зв'язку з цим, остаточні результати моделі ми будемо рахувати на базі коефіцієнту кореляції Спірмена, результати проміжних гіпотез ми подамо в обох видах.

Метод

Запропонований в статті метод є логічним продовженням експериментів описаних в [11]. Результати наших досліджень говорять про те, що ступінь зв'язності слів дійсно може бути обчислено на достатньо великому текстовому корпусі, причому корпус має бути репрезентативним, тобто відображати реалії мови без особливих викривлень. Маючи такий корпус, для підрахунку семантичної зв'язності потрібно обрати модель дистрибутивної семантики, що максимально точно зможе відобразити реалії мови. Коли таку модель знайдено залишається підібрати правильні параметри. Вибір такої моделі і її параметрів є надзвичайно складною проблемою, котру до сих пір було

вирішено значно гірше, ніж аналогічну проблему для семантичної схожості (в випадку семантичної близькості повідомлена максимальна точність на WS353-R Spearman's rho=0.81 [8], для зв’язності максимально повідомлена точність на WS353-R Spearman's rho=0.70 [12]).

В основі нашої моделі лежить гіпотеза про наявність кореляції між ступенем зв’язності двох слів та їх розподілом в корпусі. Ключовим поняттям для визначення ступеню кореляції (а значить і зв’язності) є контекст. Очевидно, що ступінь зв’язності понять залежить від частоти їх сумісних входжень в один контекст. Визначенню типу даної залежності і її математичній формалізації і присвячені всі дослідження семантичної зв’язності.

Залежність зв’язності двох понять від контексту зображено на рис. 1. Дуги демонструють взаємне входження слів в одне речення і кількість таких входжень.

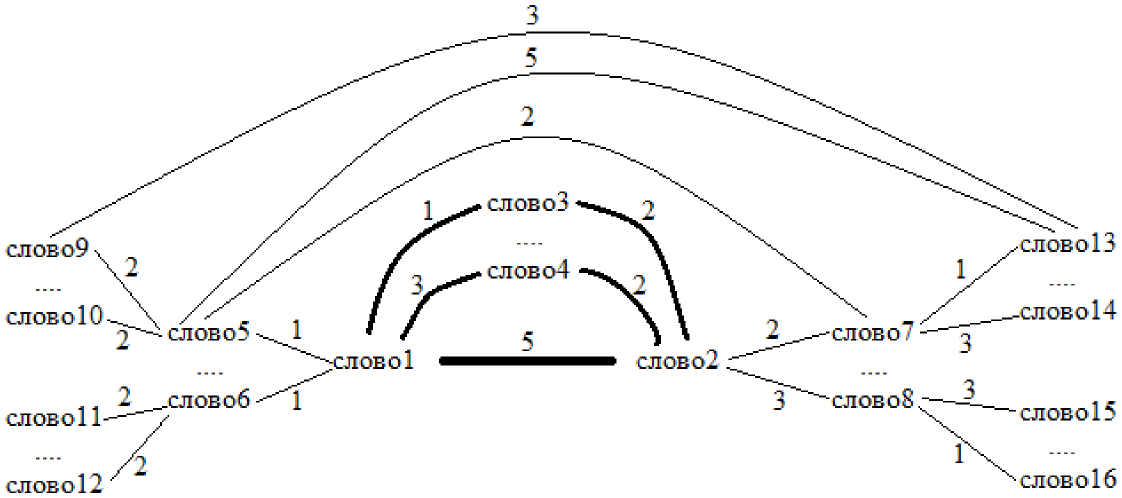


Рис. 1. Графік залежностей між словом1 і словом2

Як видно з малюнку, будь-яка пара понять, крім прямої залежності, має велику кількість залежностей через другий, третій та інші рівні понять. Визначити точну формулу впливу кожного рівня на результуючу зв’язність понять є складною задачею, яку на сьогоднішній день до кінця не вирішено. До того ж не є до кінця зрозумілим характер впливу різних даних на результат: лінійний, логарифмічний, експотенційний чи інший.

Іншим складним моментом є питання отримання однакових або хоча б близьких результатів на різних корпусах. Відмінність корпусів легко продемонструвати в термінах рис. 1. В іншому корпусі на малюнку може бути повністю відсутнє піддерево з вершиною в «слово7», ваги в піддереві з вершиною «слово1» можуть суттєво відрізнятися, і «слово1» може мати прямий зв’язок з «слово8».

Додаткова складність існує на моменті оцінки якості отриманого методу. Справа в тому, що всі чотири корпуси, що застосовуються для оцінки точності систем, було створено різними людьми. В залежності від інформаційного середовища, в якому живе людина, особливостей суспільства та професії зв’язність між поняттями може відрізнятися. Всі корпуси з еталонними значеннями формуються як середнє значення між всіма оцінками експертів, тобто, навіть в середині однієї команди оцінки можуть суттєво відрізнятися. Між корпусами різниця може бути ще більш істотною, особливо, якщо корпуси створено в різних країнах.

Для вирішення наведених питань ми пропонуємо використовувати комбінований статистично-машинний метод. Метод ґрунтується на зборі статистичної інформації з

тексту, і створенні гіпотез, що відображають різні характеристики зв'язності понять. На даний момент нами створено більше 70 гіпотез щодо визначення зв'язності. Після попереднього відбору, найбільш перспективні гіпотези передаються в модель машинного навчання.

Навчання моделі для кожного еталонного корпусу відбувається окремо. Кожен корпус ділиться на трейн та тест набори. Трейн використовується на етапі навчання для підбору коефіцієнтів під даний еталонний корпус, для кожної пари понять з навчальної вибірки, модель на вхід приймає вектор обчислених гіпотез і оцінку зв'язності цієї пари з еталонного корпусу. Після того, як тренування моделі закінчено відбувається її оцінка на тест наборі – всі пари понять подаються їй на вхід, обчислюється вектор гіпотез, прогнозується значення зв'язності. Спрогнозовані значення зв'язності формують набір оцінок, який потім порівнюється з еталонним набором за двома описаними вище метриками кореляції.

Реалізація

Для підрахунку зв'язності за запропонованим методом необхідно вираховувати велику кількість статистик з корпусу. Оскільки ступінь зв'язності понять розраховується на базі контекстів цих понять, а збір контекстів з великого текстового корпусу є надзвичайно обчислювально інтенсивною операцією, то для прискорення експериментів на етапі передобробки ми зібрали контексти всіх понять з корпусу.

Кожному поняттю відповідає один «контекст» – файл, що містить всі речення корпусу, в яких хоча б раз зустрічається дане поняття. Файл містить речення з корпусу зі спеціальною розміткою: розмічені частини мови, іменовані сутності, стоп-слова замінено спеціальною міткою STOPWORD, короткі слова – SHORT, числа замінено міткою DIGIT і т.д. Всі перетворення виконано з метою пришвидшення подальшого аналізу контексту без виклику лінгвістичних модулів. Приклад розміченого речення:

John_NE Lackey_NE DIGIT hold_VBD STOPWORD Orioles_NE hitless_NN STOPWORD Adam_NE Jones_NE homer_VBD STOPWORD one_CD STOPWORD STOPWORD STOPWORD seventh_JJ STOPWORD DIGIT STOPWORD season_NN

Коли нам необхідно обрахувати гіпотези зв'язності між поняттями t1 і t2, ми збираємо статистики з файлів context(t1) і context(t2), потім збираємо статистики зв'язків другого рівня: для множини понять T1_2={t*|t* context(t1)}, відкриваємо набір контекстів C1_2={context(t*)|t* T1_2} і з кожного з них отримуємо інформацію.

Як описано вище, ми обраховуємо більше 70 гіпотез на базі контексту. Важлива ремарка: більшість статистик сумісного входження терміна в контекст не є симетричною. Наприклад: count(t2 in context(t1))=1, в той час як count(t1 in context(t2))=2. Таке викривлення пов'язано з частим дублюванням терміну в реченні. Приклад наведений вище демонструється реченням:

word1 word2 t1 t2 word3 word4 t1.

Всі гіпотези характеризують різні риси сумісного входження термінів. Тобто, може бути гіпотеза «відношення частоти входження поняття t1 в контекст поняття t2 до частоти входження поняття t2 в контекст поняття t1», більш формально: (1)

Проте, не може існувати гіпотези «середня довжина контексту для поняття t1». Через несиметричність статистик сумісного входження, більшість гіпотез існує в трьох видах:

1. статистики пораховані на context(t1)
2. статистики пораховані на context(t2)

3. середнє перших двох статистик

Другий та третій пункти даної класифікації ми називаємо підгіпотезами.

Опишемо основні класи використовуваних гіпотез, для простоти сприйняття, поняття t1 позначимо як A, а поняття t2 як B. Для скорочення опису ми будемо застосовувати неформальну нотацію підрахунку значень гіпотез.

1. Базові гіпотези – відображають частотні характеристики сумісного входження.

1.1. R(A,B) = count(B in context(A))/count(most frequent noun in context(A))

1.1.1. Підгіпотеза1: R(B,A) = count(A in context(B))/count(most frequent noun in context(B))

1.1.2. Підгіпотеза2: Ravg(A,B) = R(A,B) + R(B,A)

1.2. R(A,B) = count(B in context(A))/count(A in context(A))

1.3. R(A,B) = count(B in context(A))/count(3rd-most frequent noun in context(A))

1.4. R(A,B) = count(B in context(A))

1.5. інші

2. Гіпотези з нормалізацією за довжиною контексту/контекстів.

2.1. R(A,B) = count(B in context(A))/(sentences in context(A))

2.2. R(A,B) = count(B in context(A))/(docs in context(A))

2.3. R(A,B) = count(B in context(A))/((docs in context(A))*(sent in context(A)))

2.4. Інші

3. Гіпотези з нормалізацією за кількістю слів.

3.1. R(A,B) = count(B in context(A))/count(unique nouns in context(A))

3.2. R(A,B) = count(B in context(A))/count(nouns in context(A))

3.3. R(A,B) = count(B in context(A))/count(unique normal forms in context(A))

3.4. R(A,B) = count(B in context(A))/count(normal forms in context(A))

3.5. R(A,B) = count(B in context(A))/(count(normal forms in context(A))+count(special markup in context(A)))

3.6. Інші

4. Гіпотези на базі відстаней.

4.1. R(A,B) = SUM(ABS(Distance(between A and B in each sentence of context(A))))

4.2. R(A,B) = 1/SUM(ABS(Distance(between A and B in each sentence of context(A))))

4.3. R(A,B) = 1/SUM(Distance(between A and B in each sentence of context(A)))

4.4. R(A,B) = 1/SUM(ABS(Distance(between A and B)/(sentence length) for each sentence of context(A))))

4.5. R(A,B) = SUM(ABS(Distance(between A and B in each sentence of context(A)))/count sentences in context(A))

4.6. Інші

5. Гіпотези з нормалізацією на базі кількості документів.

5.1. R(A,B) = (count(B in context(A))+count(A in context(B)))/((total docs in context(A)) + (total docs in context(B))

5.2. Аналогічно до 5.1, але відрізняється підгіпотеза1

5.2.1. R(A,B) = 1/SUM(ABS(Distance(between A and B)/(sentence length) for each sentence of context(A))))

5.3. Інші

6. Гіпотези на базі підрахунку зважених відстаней у графі.

7. Набір методів з варіативним обчисленням комбінованої інформації.

8. Набір методів з логарифмуванням значень та варіативним обчисленням комбінованої інформації.

9. Клас гіпотез на базі модифікації PMI(pointwise mutual information)[13]

10. Гіпотези, що обраховують зв'язність тільки для слів з сумісним входженням вище набору порогових значень та змішані гіпотези на базі PMI, обчислені над статистиками з інших гіпотез.

В пункті 1.1 ми показали як виглядають підгіпотези, оскільки методика їх формування стандартна для всіх гіпотез, то в інших випадках ми явно не вказуємо підгіпотези. На базі даних гіпотез з підгіпотезами створюються набори для навчання машинної моделі, як це викладено в описі методу.

Методика випробувань

Випробування проводились на корпусі текстів новин, зібраних на протязі 2013 року з різних новинних ресурсів: AP[14], Noodls[15], PR News Group[16] та інших. В цілому, корпус мультитематичний з деяким зміщенням в бік бізнесових та економічних новин. Всього зібраний корпус містить близько 186тис. новин, після видалення неангломовних статей, коротких статей (до 400 символів) та неповних статей в корпусі лишилося 183,420 англомовних статей придатних для обробки. В середньому розмір однієї статті в корпусі коливається від 3 до 50 Кбайт. На корпусі було зібрано контексти для всіх понять, що зустрічалися в не менше ніж шести документах.

На першому етапі побудови моделі з множини всіх гіпотез обирається підмножина. Відбір проводиться жадібним алгоритмом або Lasso regression [17]. На даній підмножині гіпотез проводиться тренування моделі на навчальний вибірці кожного з еталонних корпусів. В результаті отримуємо чотири навчених моделі – по одній для кожного еталонного корпусу. Після цього проводиться оцінка моделі на тестовій вибірці, вираховуються коефіцієнти кореляції Спірмена та Пірсона. Зрозуміло, що не жадібний алгоритм, ні Lasso не гарантують знаходження глобального оптимуму і визначені даними методами оптимальні підмножини гіпотез можуть відрізнятися, в залежності від порядку слідування гіпотез, початкового набору гіпотез, параметрів моделі, даних що ввійшли в навчальну вибірку та інших параметрів. Результати роботи моделі на різних наборах гіпотез будуть відрізнятися. Найкращі з отриманих на даний момент результатів моделі, вказано в наступному пункті.

Результати

Кожна з описаних вище гіпотез є підходом до визначення семантичної зв'язності і має право на застосування самостійно, окремо від інших гіпотез. Запропонована модель є комбінацією підходів з інтелектуально підібраними коефіцієнтами, що відображають особливості інформаційного середовища, притаманного авторам еталонного корпусу. Наведемо спочатку отримані коефіцієнти кореляції для кожної зі створених гіпотез.

На рис. 2-5 вертикальна вісь показує степінь кореляції і приймає значення в межах [-1,1], горизонтальна показує номер гіпотези в межах [1,180]. Внизу графіків показано до якої групи гіпотез належить кожна зона. З графіків видно, що кореляція значень однієї гіпотези з різними еталонними корпусами може суттєво відрізнятися, що підтверджує існування залежності між ступенем зв'язності понять в еталонному корпусі та інформаційним середовищем, в якому живуть його автори.

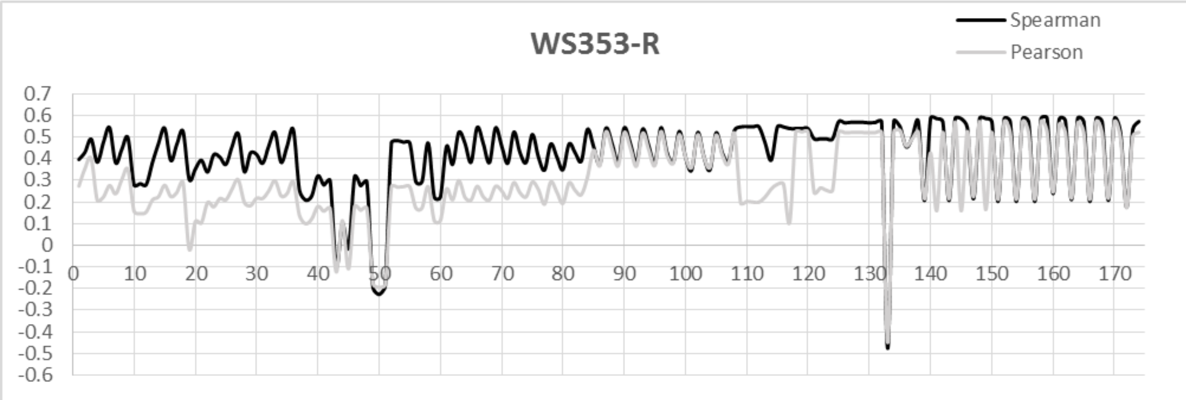


Рис. 2. Графік кореляції гіпотез з еталонним корпусом WS353-R

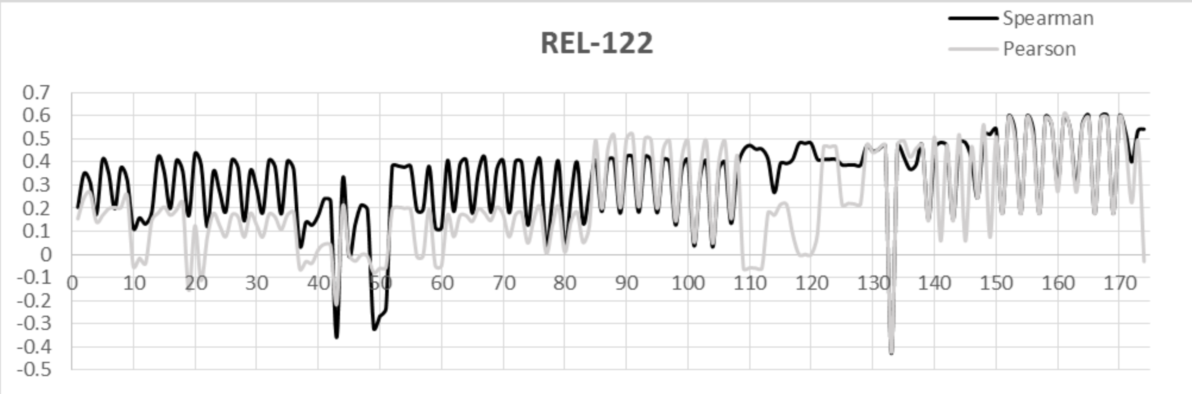


Рис. 3. Графік кореляції гіпотез з еталонним корпусом REL-122

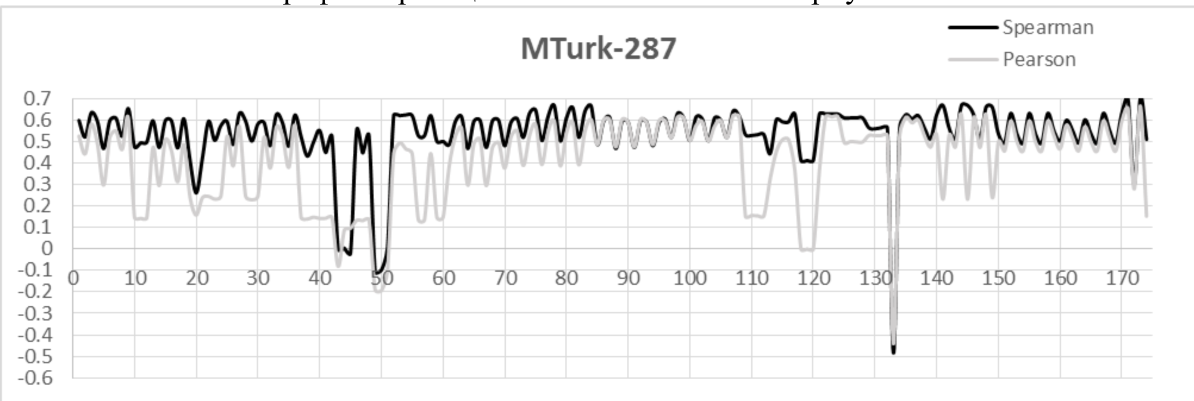


Рис. 4. Графік кореляції гіпотез з еталонним корпусом MTurk-287

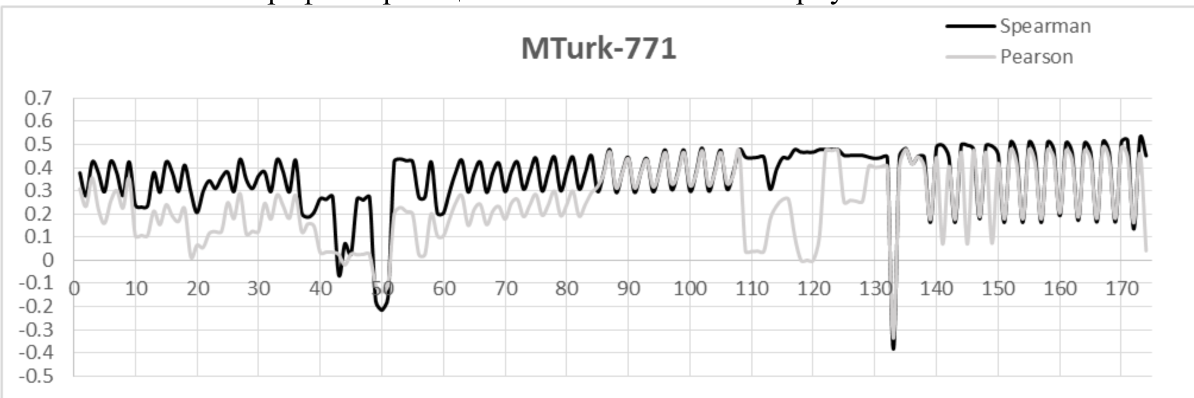


Рис. 5. Графік кореляції гіпотез з еталонним корпусом MTurk-771

Певні висновки можна зробити і про поведінку коефіцієнтів кореляції в рамках визначення зв'язності. В середньому, на однакових гіпотезах коефіцієнт Спірмена приймає вищі значення, порівняно з коефіцієнтом Пірсона. Також видно, що для більшості еталонних корпусів гіпотези в середньому приймають значення в діапазоні [0.4, 0.6]. Використання їх сукупності в запропонованій моделі дозволяє суттєво підвищити результат.

Серед множини навчених на описаних гіпотезах моделей було отримано наступні результати (наведено результати кращої моделі по кожному з еталонних корпусів). Результати наведено разом з результатами інших існуючих систем для порівняння. Через причини, описані вище, стандартом для оголошення досягнутого рівня кореляції між методом та еталонним корпусом є коефіцієнт кореляції Спірмена, тож для оцінки отриманих результатів, ми теж будемо його використовувати.

Таблиця 1. Оцінки Спірмена (ρ) на еталонних корпусах для різних методів визначення семантичної зв'язності.

	WS353-R	REL-122	MTurk-287	MTurk-771
	ρ	ρ	ρ	ρ
LG/50/m/S [18]	0.589	0.537		
PPMI [19]	0.678		0.659	
Singular [19]	0.684		0.581	
PSD-25K [19]	0.676		0.678	
DEA [20]	0.700		0.710	
Szumlanski and Gomez [21]		0.534		
CLEAR[22]				0.720
APRM avg (best)[22]			0.650	0.660
CLEAR + TSA[23]			0.751	0.742
WTMGW[24]				0.480
Multifeature models (наші моделі)	**0.745**	**0.670**	**0.770**	**0.780**

Висновки

В даній роботі нами запропоновано новий метод вимірювання семантичної зв'язності. В якості бази для вимірювання виступає великий текстовий корпус, а в якості засобу вимірювання запропоновано велику кількість гіпотез, що відображають різні аспекти дистрибутивної семантики текстів. Деякі з гіпотез самі по собі, без застосування моделі, показали високу кореляцію з еталонними корпусами, наприклад, гіпотези 9 і 10 групи на REL-122 показали $\rho>0.6$, що є показником вищим за state-of-the-art для даного корпусу. На MTurk-287 гіпотези 6-ї, 8-ї та 10-ї груп показали $\rho\approx0.7$, що є близьким до state-of-the-art. Використання множини гіпотез в рамках запропонованої моделі дозволило значно покращити результати окремих гіпотез, а застосування вагових коефіцієнтів дозволило індивідуально модифікувати важливість гіпотез для кожного еталонного корпусу.

Ми продовжуємо роботи з пошуку оптимальної підмножини гіпотез для кожного з еталонних корпусів, проте вже зараз можна стверджувати, що запропонована модель є ефективним засобом для визначення семантичної зв'язності.

Література

1. Budanitsky A. Evaluating wordnet-based measures of lexical semantic relatedness/ A. Budanitsky, G. Hirst - Computational Linguistics, 32(1), 2006 - pp. 13–47.
2. Miller G.A. WordNet: A Lexical Database for English/ G.A. Miller - Communications of the ACM, Vol. 38, No. 11, 1995 – pp. 39-41.
3. Lenat D.B. CYC: a large-scale investment in knowledge infrastructure/ D.B. Lenat - Communications of the ACM, Vol. 38, No. 11, 1995 - pp. 33-38.
4. Морозова Ю. И. Извлечение переводного словаря значимых словосочетаний из параллельных текстов с использованием методов дистрибутивной семантики/ Морозова Ю.И.// Новые информационные технологии в автоматизированных системах: материалы шестнадцатого научно-практического семинара. - М.: Моск. ин-т. электроники и математики национального исследовательского университета «Высшая школа экономики», 2013 - С. 268-272.
5. Agirre E. A study on similarity and relatedness using distributional and WordNet-based approaches/ E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa// Ann. Conf. of the North American Chapter - the Association for Computational Linguistics, 2009 – pp. 87-95.
6. Szumlanski S. A New Set of Norms for Semantic Relatedness Measures/ S. Szumlanski, F. Gomez, and V. Sims. - ACL '13, 2013 - pp. 890—895.
7. Radinsky K. A word at a time: computing word relatedness using temporal semantic analysis/ K. Radinsky, E. Agichtein, E. Gabrilovich, S. Markovitch// Proceedings of the 20th international conference on World wide web, Hyderabad, India, 2011 – pp. 172-180.
8. Guy H. Large-scale learning of word relatedness with constraints/ H. Guy, G. Dror, E. Gabrilovich, and Y. Koren // Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2012 - pp. 1406-1414.
9. Spearman C. The proof and measurement of association between two things/ Spearman C. - American Journal of Psychology N15, 1904 – pp. 72–101.
10. Rodgers J.L. Thirteen ways to look at the correlation coefficient/ J.L. Rodgers, W.A. Nicewander - The American Statistician, 42(1), 1988 – pp. 59-66.
11. Никоненко А.О. Дослідження статистичної схожості-зв’язності/ Никоненко А.О. // Вісник КНУ імені Тараса Шевченка, серія фізико-математичні науки. — 2016. — № 1 — С. 131—136.
12. Baroni M. Don’t count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors/ M. Baroni, G. Dinu, G. Kruszewski - In ACL, 2014 – pp. 238–247.
13. Rosenfeld R. A maximum entropy approach to adaptive statistical language modeling computer speech and language/ R. Rosenfeld - Computer Speech and Language, 10, 1996 – pp.187–228.
14. Associated Press [Електронний ресурс]. – Режим доступу: http://www.ap.org/
15. Gateway to facts [Електронний ресурс]. – Режим доступу: https://noodls.com/
16. Public Relations News [Електронний ресурс]. – Режим доступу: http://www.prnewsonline.com/
17. Tibshirani R. Regression Shrinkage and Selection via the lasso/ Tibshirani R. // Journal of the Royal Statistical Society. Series B (methodological) 58 (1). Wiley, 1996 – pp. 267–288.
18. Vilnis L. Word Representations via Gaussian Embedding/ L. Vilnis, A McCallum // International Conference on Learning Representations (ICLR), 2015 – pp. 128-136.
19. Li S. A generative word embedding model and its low rank positive semidefinite solution/ S. Li, J. Zhu, C. Miao // In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 2015 - pp. 1599–1609.
20. Cai Y. Differential Evolutionary Algorithm Based on Multiple Vector Metrics for Semantic Similarity Assessment in Continuous Vector Space/ Y. Cai, W. Lu, X. Che, K. Shi - DMS 2015 – pp. 241-249.
21. Szumlanski S. Automatically acquiring a semantic network of related concepts/ Szumlanski S., Gomez F. // In Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM), 2010 – pp. 19–28.
22. Jabeen S. Exploiting Wikipedia semantics for computing word associations / Jabeen S., Victoria University of Wellington, 2014 – pp.54-62.
23. Halawi G. Large-scale learning of word relatedness with constraints/ G. Halawi, G. Dror, E. Gabrilovich, Y. Koren // In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD’12, New York, USA, 2012 - pp. 1406-1415.
24. Liu B. Computing semantic relatedness using a word-text mutual guidance model/ Liu B., Feng J., Liu M., Liu F., Wang X., Li P. // NLPCC 2014. CCIS, vol. 496, Springer, Heidelberg, 2014 - pp. 67–78.

Literatura

1. Budanitsky A. Evaluating wordnet-based measures of lexical semantic relatedness/ A. Budanitsky, G. Hirst - Computational Linguistics, 32(1), 2006 - pp. 13–47.
2. Miller G.A. WordNet: A Lexical Database for English/ G.A. Miller - Communications of the ACM, Vol. 38, No. 11, 1995 – pp. 39-41.
3. Lenat D.B. CYC: a large-scale investment in knowledge infrastructure/ D.B. Lenat - Communications of the ACM, Vol. 38, No. 11, 1995 - pp. 33-38.
4. Morozova Yu.I. Izvlechenie perevodnogo slovarya znachimyih slovosochetaniy iz parallelnyih tekstov s ispolzovaniem metodov distributivnoy semantiki/ Morozova Yu.I.// Novyie informatsionnyie tehnologii v avtomatizirovannyih sistemah: materialyi shestnadtsatogo nauchno-prakticheskogo seminara. - M.: Mosk. in-t. elektroniki i matematiki natsionalnogo issledovatelskogo universiteta «Vyisshaya shkola ekonomiki», 2013 - S. 268-272.
5. Agirre E. A study on similarity and relatedness using distributional and WordNet-based approaches/ E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa// Ann. Conf. of the North American Chapter - the Association for Computational Linguistics, 2009 – pp. 87-95.
6. Szumlanski S. A New Set of Norms for Semantic Relatedness Measures/ S. Szumlanski, F. Gomez, and V. Sims. - ACL '13, 2013 - pp. 890—895.
7. Radinsky K. A word at a time: computing word relatedness using temporal semantic analysis/ K. Radinsky, E. Agichtein, E. Gabrilovich, S. Markovitch// Proceedings of the 20th international conference on World wide web, Hyderabad, India, 2011 – pp. 172-180.
8. Guy H. Large-scale learning of word relatedness with constraints/ H. Guy, G. Dror, E. Gabrilovich, and Y. Koren // Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2012 - pp. 1406-1414.
9. Spearman C. The proof and measurement of association between two things/ Spearman C. - American Journal of Psychology N15, 1904 – pp. 72–101.
10. Rodgers J.L. Thirteen ways to look at the correlation coefficient/ J.L. Rodgers, W.A. Nicewander - The American Statistician, 42(1), 1988 – pp. 59-66.
11. Nykonenko A.O., DoslIdzhennya statistichnoYi shozhostI-zv'yaznostI/ Nikonenko A.O. // VIsnik KNU ImenI Tarasa Shevchenka, serIya fIziko-matematichnI nauki. — 2016. — # 1 — C. 131—136.
12. Baroni M. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors/ M. Baroni, G. Dinu, G. Kruszewski - In ACL, 2014 – pp. 238–247.
13. Rosenfeld R. A maximum entropy approach to adaptive statistical language modeling computer speech and language/ R. Rosenfeld - Computer Speech and Language, 10, 1996 – pp.187–228.
14. Associated Press [Електронний ресурс]. – Режим доступу: http://www.ap.org/
15. Gateway to facts [Електронний ресурс]. – Режим доступу: https://noodls.com/
16. Public Relations News [Електронний ресурс]. – Режим доступу: http://www.prnewsonline.com/
17. Tibshirani R. Regression Shrinkage and Selection via the lasso/ Tibshirani R. // Journal of the Royal Statistical Society. Series B (methodological) 58 (1). Wiley, 1996 – pp. 267–288.
18. Vilnis L. Word Representations via Gaussian Embedding/ L. Vilnis, A McCallum // International Conference on Learning Representations (ICLR), 2015 – pp. 128-136.
19. Li S. A generative word embedding model and its low rank positive semidefinite solution/ S. Li, J. Zhu, C. Miao // In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 2015 - pp. 1599–1609.
20. Cai Y. Differential Evolutionary Algorithm Based on Multiple Vector Metrics for Semantic Similarity Assessment in Continuous Vector Space/ Y. Cai, W. Lu, X. Che, K. Shi - DMS 2015 – pp. 241-249.
21. Szumlanski S. Automatically acquiring a semantic network of related concepts/ Szumlanski S., Gomez F. // In Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM), 2010 – pp. 19–28.
22. Jabeen S. Exploiting Wikipedia semantics for computing word associations / Jabeen S., Victoria University of Wellington, 2014 – pp.54-62.
23. Halawi G. Large-scale learning of word relatedness with constraints/ G. Halawi, G. Dror, E. Gabrilovich, Y. Koren // In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD'12, New York, USA, 2012 - pp. 1406-1415.
24. Liu B. Computing semantic relatedness using a word-text mutual guidance model/ Liu B., Feng J., Liu M., Liu F., Wang X., Li P. // NLPCC 2014. CCIS, vol. 496, Springer, Heidelberg, 2014 - pp. 67–78.

RESUME

A.O. Nykonenko

Semantic relatedness calculation method

The article describes a new method for determination of semantic relatedness. The method is based on statistical data collected from text corpora and principles of distributive semantics. A set of basic hypotheses lies at the basis of the method. Each hypothesis is a feature of semantic relatedness itself and can be used separately from the method. Total offered more than 170 hypotheses (including sub-hypothesis). The main hypotheses can be split on the following classes:

1. Basic hypotheses - reflects the frequency characteristics of the common occurrences of the words.
2. Hypotheses with normalization by length of the context.
3. Hypotheses with normalization by the number of words.
4. Distances based hypotheses.
5. Hypotheses with normalization by the number of documents.
6. Hypotheses based on the calculation of weighted distances on a graph
7. A set of methods with variational calculation of the combined information.
8. A set of methods with logarithm of values and variational calculation of the combined information.
9. Hypotheses with different PMI modefications.
10. Hypotheses that calculate relatedness only for words with common occurrences above the certain thresholds and mixed hypotheses based on PMI, calculated over statistics from other hypotheses.

The article shows graphs for Spearman and Pearson correlations for each hypotheses on the benchmark sets. The plots contain marked boundaries for hypothesis’s classes and a correlation for each class. Also, noted various behavior of the same hypothesis on different benchmark sets, which confirms our observation of the different nature of the benchmark sets.

Further, on the basis of the proposed hypotheses, created aggregating model for evaluating semantic relatedness. Model measured on all benchmark sets. Received ratings exceeded the evaluation of other existing methods, this confirms the effectiveness of the proposed model.

Надійшла до редакції 18.10.2016

